A Suggested Model for the Origin of Mosaic Structure of DNA

Carla Goldman and Robson Francisco de Souza
Departamento de Física Geral,
Instituto de Física
Universidade de São Paulo, Caixa Postal 66318
05389-970 São Paulo, SP, Brazil
e-mail: carla@gibbs.if.usp.br

September, 1996

Abstract

In view of the critical behavior exhibited by a statistical model describing a one dimensional crystal containing isotopic impurities and the fact that for impurity mass parameter 3 times the host mass the critical temperature can attain very high values, we suggest that the mosaic structure displayed by certain nucleic acids (DNA) of eukaryotic organisms might have had its origins in a phenomenon of condensation of *codons* ocurred at prebiotic conditions. An appropriated map of that model onto some of the DNA features allows one to predict power law behavior for both correlation functions and exons size distribution in binary sequences of nucleotides which are distinguished by their protein coding or noncoding functions. Preliminary studies of these quantities performed on intron-containing sequences from *GenBank* are presented here.

1 Introduction

Nucleotide sequences on DNA of high organisms display a mosaic structure in such a way that expressed sequences (*exons*) are interrupted by intervening sequences (*introns*) which do not code for aminoacids [1]. Up to the present, the role of introns are not unraveled although there are beliefs that their presence on the eukaryotic genome promote biological stability since mutations on noncoding regions are not expected to affect information contents regarding protein synthesis [2].

Uncertainties on whether introns have been either introduced into or withdrawn from embryonic sequences of nucleotides have raised doubts on the common assumption that procaryotic genome is primitive relatively to the eukaryotic genome. Up till now, the dispute between introns-early [3] and introns-late [4] hypotheses appears to be unresolved.

It is worth to notice in this respect that there are evidences favoring the presence of exons of all sizes on eukaryotic sequences and the existence of a close relationship between exons distribution along such sequences and protein structure [5]. The question of whether such a distribution has a prebiotic origin is then closely related to the introns-early/introns-late dilemma [6].

In the last few years there have been numerous efforts to rationalize the genetic information stored in DNA nucleotide sequences from a statistical viewpoint [7]. Apart from differences of methodology in using *GenBank* data base, these works present numerical information on the two-point autocorrelation function $C(r)$ in one-dimensional space (or its Fourier transform power spectrum), for nucleotide separation distance r (in appropriate units) on a same DNA strand. The essential question addressed in these works relates to disclose long-range correlations (or, equivalently, power-law behavior) in these sequences by distinguishing nucleotides according to their pyrimidine (Cytosine, Thymine) or purine (Adenine , Guanine) base content . Criteria based on the number of hydrogen-bonds linking complementary nucleotides in pairs have also been used.

In this context, the relevance of the coding and noncoding parts of eukaryotic sequences for interpreting the genetic information of DNA has been upraised since S. Nee's' suggestion that the ultimate source of long-range correlations reported in early studies [7], could be the underlined mosaic structure of genome [8]. This observation has led to further studies on purine and pyrimidine organization along separate coding and noncoding sequences. Nevertheless, the question on whether the mosaic structure of the genome is responsible for signaling long range order in purine-pyrimidine distribution, remains under debate [9] .

From a physical viewpoint one of the interests of studying statistical properties of the genome resides on the fact that, by finding long- range behavior for $C(r)$, one might address the origin of nucleotide base sequences organization to a critical phenomenon that eventually occurred during biological evolution. In this respect, one can find in the literature nonequilibrium statistical mechanics models proposed to describe the dynamics of nucleotide sequences under the action of some natural evolutionary driving processes [10].

We choose to address questions related to the genome organization by distinguishing nucleotides not between purines and pyrimidines but instead, regardless their base content, by distinguishing nucleotides according their loci on an exon or intron region of a DNA sequence . We believe that direct information on statistical properties of mosaic structure shall be helpful for tracing the origins and functions of introns, which is our main concern here.

The focus of our studies on nucleotide sequences was entirely suggested by the results obtained for equilibrium properties of a statistical mechanics model proposed recently to investigate conditions for isotopic order (or "isotopic fractionation") on a harmonic crystal chain containing isotopic impurities [11]. On the basis of an appropriated map (see below), we argue that this model is also suitable for describing some features of DNA concerning, in particular, its mosaic structure.

The relevant aspect of the original model equilibrium thermodynamics we want to emphasize here is that at temperatures below a certain critical temperature T_C, it displays a condensed phase in which impurities aggregate into clusters of all sizes. Within this phase, the cluster distribution function and two-point auto-correlation functions are expected to exhibit temperature dependent power-law decay [12]. Moreover, it was also found in Ref. [11] that for masses m_a of host particles and m_b of impurities satisfying the particular relationship

$$m_b \sim 3m_a \tag{1}$$

the critical temperature for condensation can attain very high values [13].

Making a parallel of this model with a similar description for DNA molecule, allows us to conjecture that the origin of DNA mosaic structure might have been due to a *codon* [14] condensation into clusters (or *exons*) that eventually occurred during prebiotic synthesis of some nucleic acids. Within this picture, we suggest that thermodynamical stability of DNA with respect to small spatial fluctuations of its molecular components had a crucial role for establishment of an ordered state (mosaic structure) in which codons appear segregated from protein noncoding regions. Important to stress is that the present proposal do not relate to the emergence of particular sequences of purine and pyrimidine bases; it is restricted to the mosaic features, regardless base content.

In Section 2 we review the model of Ref. [11] and extend the results in order to estimate the asymptotic behavior of related pair correlations functions and clusters size distribution. On the basis of a map of this model onto DNA nucleotide sequences suggested in Section 3, we show some numerical results for these quantities obtained in preliminary studies using data from *GenBank*. A discussion is in Section 4.

2 A model Hamiltonian

Here we review the original model of Ref. [11] and explore some of its statistical properties to describe aspects of systems of interest.

2.1 Isotopic order : phonon induced interactions

As mentioned, the original lattice model Hamiltonian describes a one dimensional crystal containing isotopic impurities interacting by harmonic potential

$$H = \sum_{i=1}^{N} \{ \frac{m_i}{2} \dot{u}_i^{2} + \frac{K}{2} u_i (u_i - u_{i-1}) \} \qquad (2)$$

where u_i and m_i are respectively, the displacement from equilibrium position and mass of the particle at lattice site i. K is a common force constant. There are present in this system two isotopic species, A (host) and B (impurity), with masses m_a and m_b, respectively. By introducing site-dependent spinlike variables σ_i which assume the value 0 (for host species) and 1 (for impurity), it is possible to write m_i as:

$$m_i = m_a + (m_b - m_a)\sigma_i \tag{3}$$

Now, it is assumed that this system can be driven into a region of sufficiently high temperatures where the particles would freely interchange positions with each other. The nature of such processes is immaterial here, conditioned that their characteristic times are very small if compared to typical lattice relaxation times. The question posed in Ref. [11] is whether in this situation the phonons of the lattice could induce positional correlations among impurities leading to some sort of isotopic order at thermodynamical equilibrium. For that, we introduce phonon creation $\left(b_q^\dagger\right)$and phonon annihilation (b_q)operators as follows:

$$\dot{u}_j= \sum \left[\frac{1}{2Nm_a\omega_q}\right]^{\frac{1}{2}} \left(b_q + b_{-q}^\dagger\right) e^{iqj} \tag{4}$$

and

$$u_j = \sum \left[\frac{1}{2Nm_a\omega_q}\right]^{\frac{1}{2}} \left(b_q - b_{-q}^\dagger\right) e^{iqj} \tag{5}$$

rewrite H in (2) as

$$H = \sum_q \omega_q b_q^\dagger b_q + \frac{\alpha}{N} \sum_{qq'} \widetilde{\sigma}_{q-q'} (\omega_q \omega_{q'})^{\frac{1}{2}} B_q B_{q'} \tag{6}$$

where $B_q = b_q - b_{-q}^\dagger$,

$$\alpha = \frac{1}{4}\left(m_b/m_a - 1\right), \tag{7}$$

$$\widetilde{\sigma}_k \equiv \sum_j \sigma_j e^{ikj} \tag{8}$$

and $\omega_q \;\; = \left[2K(1-\cos q)/m_a\right]^{1/2}$ is the dispersion relation for free phonons for

$$q \in \{-\pi, (\frac{-N-2}{N})\pi, ..., \frac{-2\pi}{N}, 0, \frac{2\pi}{N}, ...\pi\}.$$

To approach equilibrium properties of this system, we consider its grand partition function at fixed temperature β^{-1} :

$$\Xi(\beta,\mu) = \sum_{\{\sigma_j\}} \left(\sum_{\{n_q\}} e^{-\beta H} \right) e^{\beta\mu \sum_j \sigma_j} \tag{9}$$

where the sums extend over all spins configurations $\{\sigma_j\}$ and over phonon occupation number in the q mode $\{n_q\}$. μ is a chemical potential that controls the density of impurities and H is as in (6). Integration over phonon variables allows one to write Ξ as

$$\Xi(\beta,\mu) = Z_0 \sum_{\{\sigma_j\}} e^{-\beta \Delta F(\{\sigma_j\})} \, e^{\beta\mu \sum_j \sigma_j} \tag{10}$$

where

$$\begin{aligned} \Delta F(\{\sigma_j\}) \equiv -\beta^{-1} \ln \frac{\sum_{\{n_q\}} \left\langle \{n_q\} \left| e^{-\beta H} \right| \{n_q\} \right\rangle}{\sum_{\{n_q\}} \left\langle \{n_q\} \left| e^{-\beta H_0} \right| \{n_q\} \right\rangle} \equiv \\ \equiv -\beta^{-1} \ln Z/Z_0 \end{aligned} \tag{11}$$

is an effective interaction among impurities induced then by the lattice phonon contents. Z_0 is a normalization constant defined to set $\Delta F(0) = 0$.

We've found that, for a given spin configuration and low impurity density, ΔF is separated into a sum

$$\Delta F = \sum_l \Delta F_{l^{(r)}} \tag{12}$$

such that $\Delta F_l^{(r)}$ is the effective energy of an isolated cluster of impurities (indexed by r) containing l components; namely, a l-cluster [15].

This enables us to focus on an isolated l-cluster and extract from ΔF_l a cluster "surface energy" term E_l whose asymptotic behavior is given by

$$E_l \sim \frac{(K/m_a)^{1/2}}{\pi(2-\Lambda)^2} \ln l \tag{13}$$

as l goes to infinity, with

$$\Lambda = \frac{m_b + m_a}{m_b - m_a} \tag{14}$$

and a cluster bulk energy $l\Phi$, in agreement to the model studied in Ref. [12]. Accordingly,

$$\Delta F_l \sim l\Phi + E_l \tag{15}$$

In the following we make further considerations to discuss the nature of the phase transition exhibited by this model on the basis of equations (12) and (15).

2.2 Clusters Size Distribution

It is our purpose here to characterize the phase transition displayed by the model above in the limit of very low impurity density. For this, we replace the sum over spin variables in (10) by a sum over all configurations of isolated impurity clusters of all sizes and positions. Within this approximation we treat the system as an ideal lattice gas mixture of different molecular species.

Let a_l be the number of l-clusters in a given configuration of the system. Then (12) becomes

$$\Delta F \sim \sum_{l=1}^{\infty} a_l(l\Phi + E_l) \tag{16}$$

and (10) can be approximated by

$$\Xi(\beta, \mu) \simeq Z_0 \prod_l \sum_{a_l=0}^{\infty} \exp[-\beta a_l(l\Phi + E_l)].\frac{(Lz_l)^{a_l}}{a_l!} \tag{17}$$

where L is the chain length and $z_l = e^{\beta\mu_l}$ is the l-cluster activity with μ_l being the corresponding chemical potential. At thermodynamical equilibrium,

$$\mu_l = l\mu \tag{18}$$

Defining a l-cluster *internal partition function* q_l as

$$q_l = \exp[-\beta(l\Phi + E_l)] \tag{19}$$

expression (17) reads

$$\Xi(\beta, \mu) \simeq Z_0 \prod_{l=1}^{\infty} e^{Lq_l z_l} \tag{20}$$

from which the average number of l-clusters per unit length $\rho_l(\beta, \mu)$ can readily be obtained:

$$\rho_l(\beta, \mu) = \lim_{L\to\infty} \frac{1}{L} z_l \left(\frac{\partial \ln \Xi}{\partial z_l}\right)_{L,T} = l^{-\frac{\beta}{\beta_c}}.\exp[-\beta l(\Phi - \mu)] \tag{21}$$

In expression (21), we have used relation (18) and defined a critical inverse temperature β_c by

$$\beta_c = \left[\frac{(K/m_a)^{\frac{1}{2}}}{\pi(2-\Lambda)^2}\right]^{-1} \tag{22}$$

On summing (21) over all values of l, we obtain the average number of clusters of any kind per unit length; namely,

$$\nu(\beta,\mu) = \sum_{l=1}^{\infty} \rho_l(\beta,\mu) = \sum_{l=1}^{\infty} l^{-\frac{\beta}{\beta_c}}.[e^{-\beta(\Phi-\mu)}]^l \tag{23}$$

so that the fraction of l-clusters at thermodynamical equilibrium, or equivalently, the probability of finding a l-cluster in the system, is given by

$$P_l(\beta,\mu) = \frac{\rho_l}{\nu} \tag{24}$$

Let us now review the properties of the sum in (23) [16]:
(a) If $e^{-\beta(\Phi-\mu)} < 1$, then ν converges for $\beta > 0$;
(b) If $e^{-\beta(\Phi-\mu)} = 1$, then ν converges for $\beta > \beta_c$ or $T < T_c = 1/k_B\beta_c$.

The last condition expresses the critical behavior of the model which becomes apparent by studying the limit

$$\lim_{\mu\uparrow\Phi} P_l(\beta,\mu) = \lim_{\mu\uparrow\Phi} \frac{l^{-\frac{\beta}{\beta_c}}}{\nu(\beta,\mu)}. \tag{25}$$

If $T > T_c$ then $\lim_{\mu\uparrow\Phi} \nu(\beta,\mu) \to \infty$ so that for temperatures higher than T_c the probability of finding stable clusters of relative large sizes is negligible. However, this limit attains finite values in regions where $T < T_c$. Consequently, at sufficiently low temperatures the statistical properties of the system are governed by the Lévy distribution

$$P_l(\beta > \beta_c, \Phi) \sim l^{-\frac{\beta}{\beta_c}} \equiv l^{\eta} \tag{26}$$

with

$$\eta = \eta(\beta) = -\beta/\beta_c \tag{27}$$

which characterizes its condensed phase. Within this phase macroscopic regions of impurities can be found at all scales.

We should notice that result (26) which was obtained here for noninteracting clusters, is expected to hold for this class of models even in more general cases as those studied in Ref. [12].

2.3 Pair Correlation Functions

In addition to the size distribution we can also make quantitative predictions on the behavior of impurity pair correlation functions which in the present context are defined by

$$C(|r|) = \langle \sigma_i \sigma_{i+r} \rangle - \langle \sigma_i \rangle \langle \sigma_{i+r} \rangle \tag{28}$$

for $|r|$ measuring distances between two impurity particles along the considered chain. The set of variables $\{\sigma_i\}$ are assigned according to (3). $\langle ... \rangle$ indicates averages over statistical ensemble of the system.

Since the only contribution to the truncated function (28) comes from configurations in which the considered two points are covered by a single cluster we can use the approach introduced in the last Section and estimate the behavior of $C(|r|)$ at the condensed phase by the probability of finding any cluster of size greater or equal than $|r|$. Using result (26), we then disclose asymptotic power-law behavior

$$C(|r|) \sim \sum_{l=|r|}^{\infty} P_l(\beta > \beta_c, \Phi) \sim |r|^{\varepsilon} \tag{29}$$

with

$$\varepsilon = \varepsilon(\beta) = 1 - \beta/\beta_c \tag{30}$$

as expected [12], [17]. We remark that $1 - \beta/\beta_c < 0$ within this phase.

3 Modeling Eukaryotic Sequences

3.1 Map onto DNA

It is our intention in this Section to suggest a map of the model reviewed above in order to describe aspects of DNA eukaryotic sequences. For this purpose, it is relevant to notice in Eq. (14) that for $\Lambda \sim 2$, i.e., for

$$m_b \sim 3m_a \tag{31}$$

the critical temperature of condensation T_c can attain arbitrarily large values [see Eq. (22)]. This result leads us to make some conjectures concerning coding and noncoding organization along some nucleotide sequences:

(i) Let us assume that small positional displacements of nucleotides with respect to their equilibrium positions along DNA α-helix backbone, which in turn generate phonons, are relevant degrees of freedom of the macromolecule and can be accounted by the lattice model Hamiltonian in Eq.(2). For this, we assume also that the two DNA complementary strands can be represented by a single chain of harmonic oscillators placed along that direction. Accordingly, we shall consider that any two nucleotides linked by H-bonds in complementary pairs are, in certain conditions, constrained to move along chain direction as a single particle of mass (Fig 1a):

$$m + m_A + m_T = m_{AT} \tag{32}$$

or

$$m + m_C + m_G = m_{CG} \tag{33}$$

where $m_{A(T,C,G)}$ is the mass of an Adenine (Thymine, Cytosine or Guanine) base and m is the mass of remaining sugar and phosphate components of nucleotides.

FIGURES 1(a) ,1(b)

Due to the fact that either complementary pair AT or CG comprise a two-carbon ring purine base and a one-carbon pyrimidine joint to attain a three-ring base pair (see Fig.1b), we can ascribe within reasonable approximation the mass of a host particle m_a defined in the model for isotopes, to the mass of any combination A-T or C-G, i.e.

$$m_a \sim m_{AT} \sim m_{CG} \tag{34}$$

(ii) In addition, we consider the possibility that eventually in our DNA lattice model there are present some triplets of consecutive complementary pairs of nucleotides which are also constrained (by action of mechanisms external to the system) to move as single particles; each of these comprises then six nucleotides. The number of such assembled triplets is controlled by a chemical potential μ.

In view of (34) we can then ascribe a mass m_b of "impurities" particles to any of these triplets in such a way that relation (31) holds, i.e.

$$m_b \sim c_1 m_{AT} + c_2 m_{CG} \tag{35}$$

for c_1, c_2 integers assuming any value of the interval $[0, 3]$ but subjected to the condition $c_1 + c_2 = 3$.

(iii) Finally, the assembled triplets are to be identified here with *codons* (accompanied by their complementary nucleotides), which are the units of genome contents information.

In analogy to the model of last Section, we assign site dependent spinlike variables $\{\sigma\}$ to the lattice ordered points such that

$$\begin{array}{lll} \sigma_i = 1, & \text{if} & i \text{ is a codon} \\ & \text{or} & \\ \sigma_i = 0, & \text{if} & i \text{ is not a codon} \end{array} \tag{36}$$

Notice that by this procedure, each codon occupies a single lattice site. Notice also that, according to our purposes, the defined set $\{\sigma\}$ for chosen nucleotide sequences disregard differences concerning the nature of purine-pyrimidine base pairs.

Steps (i), (ii) and (iii) accomplish the map. Let us examine some of its consequences.

3.2 Distribution of Coding Sequences: Mosaic Structure

We are now able to make some quantitative predictions on the structure of eukaryotic DNA sequences with respect to the distribution of coding regions.

According to the results of Sec. 2.2, and the above map, we expect that the size distribution of coding regions along eukaryotic gene sequences, which we refer here as DNA mosaic structure [18], follows the power-law (26) with $\beta/\beta_c > 1$, in analogy to the distribution of clusters of isotopic impurities of the physical model at condensed phase.

In order to get support for this conjecture we select sequences from *GenBank* and check for this property. We focus on Saccharomyces chromosomes (SCCHRIII, SCCHRIX, YSCCHRVIN) which are among the longest intron-containing sequences available at that data bank. We notice that such sequences were selected from available data upon requirement of presenting complete coding as well as noncoding regions (CDS sequences).

Figure 2 shows histograms where there are depicted the diverse sizes of the respective coding regions. To extract representative functions for these diagrams in each case we test both polynomial and exponential fittings which are also shown.

FIGURE 2

3.3 Codon Pair Correlation Functions

In the same context we examine the behavior of codon pair correlation functions. Figure 3 shows the variation of $C(r)$ with codon pair distance r for various intron-containing nucleotide sequences taken from *GenBank*. These curves have been obtained upon assignments (36), and assuming ergodicity of the systems of interest so that for each sequence, we replace the ensemble averages indicated in (28) by "temporal" averages

$$C(|r|) = \frac{1}{L}\sum_{i=1}^{L}\sigma_{i+r}\sigma_i - \left(\frac{1}{L}\sum_{i=1}^{L}\sigma_{i+r}\right)\left(\frac{1}{L}\sum_{i=1}^{L}\sigma_i\right) \qquad (37)$$

where the sequence length L is measured in nucleotide units.

Fig. 3 shows our results for $C(r)$ obtained by evaluation of expression (37) on the same DNA sequences studied last Section and considering periodic boundary conditions.

FIGURE 3

Comments on these results are in Sec.4.

4 Discussion

The above ideas relate to the origin and function of genome mosaic organization. They are based on the results obtained previously for equilibrium properties of a one-dimensional harmonic lattice chain model of two isotopic components (masses m_a and m_b) which was proposed recently for studying isotopic fractionation. In favorable thermodynamic conditions, this model exhibits a condensed phase in which segregation takes place, giving rise to macroscopic regions of just one component [11]. In order to use these results to pursue a description of DNA mosaic structure, we focus on the multiscale characteristic predicted for the distribution of such aggregates and on the fact that the critical temperature for phase transition can attain very high values for a particular set of values for model parameters (i.e. for $m_b \sim 3m_a$). In this case, the condensed phase should be stable in a broad range of temperatures including, in particular, our present room temperatures.

On the basis of the map of Section 3, we are then lead to suggest: (i) that the mosaic structure might have had its origins in a condensation phenomena of *codons* that took place at prebiotic conditions, and (ii) this mosaic structure, being *thermodynamically stable* with respect to the distribution of coding and noncoding sequences, might in fact be related to the *biological stability* of the genome of high organisms since it is believed that the presence of noncoding sequences ensures relative low probabilities of harmful mutations on the genome [2].

In order to find support to the model and to check for the plausibility of related assumptions, we explore the fact that both exon size distribution functions and codon correlation functions are predicted to display power-law decay according to expressions (26) and (29) respectively. It is remarkable in this respect that the decay power in both cases are predicted to be temperature dependent. This fact shall be of some relevance from a biological point of view, for it provides information on the thermodynamics of gene formation.

We select sequences from *GenBank* for which calculation of these quantities were performed. The results obtained are shown in Figures 2 and 3. In these calculations we distinguish nucleotides according to their loci on a coding or noncoding region of each sequence, in agreement to our assumptions.

Let us focus on the results obtained for pair correlations of selected chromosome sequences, and restrict the analysis to the range of two-point distances depicted in Figure 3. We notice that correlation curves for all chosen sequences are smooth only in a relative small range, exhibiting in the remaining an oscillatory behavior due, probably, to the absence of experimental information on coding and non-coding regions in all representative scales [19]. Even so, we observe that the decay profile of the curves in Figure 3 can in principle be fitted by either exponential or polynomial functions. Although the characteristic decay parameter in each of the exponentials shown is very small (i.e., not greater than O (10^{-3}), see Table 1) we should not, on the basis of these results alone, discard the possibility of exponential in favor of polynomial fittings. In face of this, the situation then appears to be inconclusive concerning long-range order of coding regions in eukaryotic sequences of nucleotides.

Sequence	m^{exp}	n^{exp}	$\varepsilon^{\mathrm{exp}}$	η^{exp}	$\varepsilon^{\mathrm{exp}} - \eta^{\mathrm{exp}}$
SCCHRIII	-1.4×10^{-3}	-2.1×10^{-3}	$-.79$	-1.83	1.04
SCCHRIX	-7.7×10^{-4}	-6.2×10^{-4}	$-.99$	-2.15	1.16
YSCCHRVIN	-8.3×10^{-4}	-2.6×10^{-3}	$-.91$	-1.68	.77

Table 1 - Results of polynomial fitting exponents for exons size distribution, η^{exp}, and codon correlation functions, $\varepsilon^{\mathrm{exp}}$ (Fig.2 and 3 respectively), along the saccharomyces chromosome nucleotide sequences indicated. Experimental information on coding and non-coding regions along each of these sequences were taken from *GenBank* data bank. m^{exp} and n^{exp} are the corresponding results for exponential fitting coefficients. Last column shows the difference $\eta^{\mathrm{exp}} - \varepsilon^{\mathrm{exp}}$ obtained in each case. According to our theoretical predictions, this difference is expected to be of order one.

Let us now focus on the numerical results obtained for exons size distribution. The fact that there are present in the histograms many gaps in different positions corroborates our conjectures above on the oscillatory behavior of correlation functions. Figure 2 presents both exponential and polynomial fittings for the referred histograms. We observe that also in this case either fittings seem to be adequate although, as for the case of correlations, the decay parameter of each of the fitted exponentials is very small.

Additional evidences in favor of polynomial fittings can nevertheless be obtained in case of distribution functions. In fact, as can be observed in Figure 2, rare events, namely, relative large coding regions present in each sequence can only be accounted by polynomial fitting curves with $\beta/\beta_c > 1$, in agreement to our predictions.

Moreover, if we compare the decay profiles of related correlation and distribution polynomial fittings we find that the differences between the corresponding decay coefficients for all of the sequences studied is of order one, as predicted by our model [compare expressions (26) and (29) with results in the last column of Table 1).

In conclusion, although our numerical studies lack more information on other scales of coding and non-coding regions of genome, it is still possible to obtain strong support for the use of the model of Section 2 to describe aspects of DNA mosaic structure, on the basis of a joint analysis of available experimental data on both exons size distribution and codon correlation functions.

Finally, we comment on the hypothesis of Section 3. In our opinion, it is important for the model credibility to find reasons at least for some of the stated assumptions, which we review here: a) in our DNA lattice model, some triplets of nucleotides complementary pairs are constrained to behave as single particles concerning their interactions with the remaining system, in contrast to other pairs which are not constrained in this way; b) in special conditions (prebiotic), all particles can interchange positions with each other.

One possibility to think on these is to imagine triplets being assembled at the required prebiotic conditions, by the action of external agents that are able to promote aggregation of any three consecutive nucleotide complementary pairs at random positions on the lattice. Details of physical mechanisms involved in this kind of process are immaterial here, since it is also assumed that it happens in a in a very short time scale if compared to the lattice relaxation time. Within this picture, the coding regions of DNA would emerge due to a possible sequential action of these external agents. According to the results presented here, stabilization of macroscopic sequences of such triplets though, would be favorable only at regions of temperatures below T_C.

5 Acknowledgments

This work was supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Brazil.

C.G. acknowledges D.H.U Marchetti for illuminating discussions and for reading the manuscript.

References

[1] - Berger et al, Proc. Natn. Acad. Sci. USA 74, 3171 (1977);

[2] - J.D. Watson, N.H. Hopkins, J.W. Roberts, J.A. Steitz, A.M. Weiner, *Molecular Biology of the Gene*, fourth edition (The Benjamin Cummings Publishing Company Inc., Ca 1987);

[3] - - W. Gilbert, Nature 271, 501 (1978);

[4] - W.F. Doolittle, Nature 272, 581 (1978);

[5] - A. Stoltzfus, D.F. Spencer, M. Zuker, J.M. Logsdom Jr., W.F. Doolittle , Science 265, 202 (1994);

[6] - L.D. Hurst, Nature 371, 381 (1994);

[7] - For earlier references on the subject, see W.Li, Europhys. Lett. 10, 395 (1989); C.K. Peng, S.V. Buldyrev, A.L. Golberger, S. Halvin, F. Scortino, M. Simons and H.E. Stanley, Nature 356, 168 (1992); R.Voss, Phys. Rev. Lett 68, 3805 (1992); B. Borstnik, D. Pumpernik, D Lukman, Europhy. Lett. 23 , 389 (1993);

[8] - S. Nee, Nature 357, 450 (1992); S. Karlin and V. Brendel, Science 259, 677 (1993); D. Larhammar and C.A. Chatzidimitrio- Dreissman, Nucl. Acids Res. 21, 5167 (1993);

[9] - S.V. Buldyrev, A.L. Goldberger, S. Havlin, C-K Peng, M. Simons, H.E. Stanley, Phys. Rev E 47, 4514 (1993); Phys. Rev. E 49, 1685 (1994); Phys. Rev. Lett. 73, 3169 (1994); Phys. Rev E 51, 5084 (1995); W.Li, T.G. Marr and K. Kaneko, Physica D 75, 392 (1994); A. Arneodo, E. Bacry, P.V. Graves and J.F. Muzy, Phys.Rev.Lett 74, 3293 (1995); P.P.Galvaán, R.R. Roldán and J.L.Oliver, Phys. Rev E 53, 5181 (1996);

[10] - These are based on Cellular Automata models. See for example, H.Gutowitz, ed. Cellular Automata: Theory and Experiment, Physica D45, 1990 (North Holland, Amsterdam, 1990); (MIT Press, Cambridge, MA, 1990);

[11] - C. Goldman and A. Berezin, Phys. Rev. B 51, 12361 (1995);

[12] - B.U. Felderhof and M.E. Fisher, Ann of Phys. 58, 176 (1970); Ann of Phys. 58, 217 (1970); Ann. of Phys. 58, 268 (1970); Ann. of Phys. 58, 281 (1970);

[13] - The region of model parameters considered in Ref.[11] was restricted to those values for which $m_b < m_a$. Extension of the results for all regions can be easily obtained and the limit of $m_b \sim 3m_a$ can be taken.

[14] - In biological literature a codon refers to a set of three consecutive nucleotides on a DNA sequence that codifies for one specific protein aminoacid.

[15] - According to Ref. [11], an isolated l-cluster in the present context is defined as a set of l impurities occupying consecutive sites of the lattice chain and separated from other impurities at least by one host site at each end.

[16] - I.S. Gradshteyn and I.M. Ryzhik, Table of Integrals, Series and Products (Academic, San Diego, 1980).

[17] - M.E. Fisher, Physics 3, 255 (1967).

[18] - In general, regions of nucleotides on an eukaryotic DNA sequence comprise additional classification (e.g., flanking, operons, etc) besides of being classified primarily as *exons* or *introns*. To this fully structured sequence one reports to as the *mosaic structure* of DNA. The simplified view adopted here introduces no ambiguities and is intended only to emphasize the multiscale characteristics of the genome with respect to the coding and non-coding regions.

[19] - The oscillatory behavior of correlation functions shall be smoothed by accounting for an eventual degree of connectivity (interaction) among the diverse coding regions along the considered sequence; such interactions shall then preserve scale invariance of the system. We should emphasize however that such a possibility has been totally disregarded in our calculation.

6 Figure Caption

Figure 1a: Cytosine-Guanine (CG) and Adenine-Thymine (AT) nucleotide complementary pairs. Notice that either CG or AT comprises three aromatic rings.

Figure 1b: Single chain mass-string representation of a DNA nucleotide sequence.

Figure 2: Exon size histograms for saccharomyces chromosome nucleotide sequences taken from *GenBank* : (a) SCCHRIII [315,341 base pairs (bp)]; (b) SCCHRIX (439,885 bp) and (c) YSCCHRVIN (270,148 bp). It is also shown polynomial $P^*(l) \sim l^{\eta^{\rm exp}}$ (solid line) and exponential $P^*(l) \sim \exp(m^{\rm exp} l)$ (dashed line) fitting curves as functions of exon size l. Results for $\eta^{\rm exp}$ and $m^{\rm exp}$ obtained from these fittings in each case are in Table1.

Figure 3: Corresponding pair correlation functions $C^*(r)$ for the sequences of Figure 2. Polynomial $C^*(r) \sim r^{\varepsilon^{\rm exp}}$ and exponential $C^*(r) \sim \exp(n^{\rm exp} r)$ fitting curves are also depicted and results for $\varepsilon^{\rm exp}$ and $n^{\rm exp}$ are in Table1.

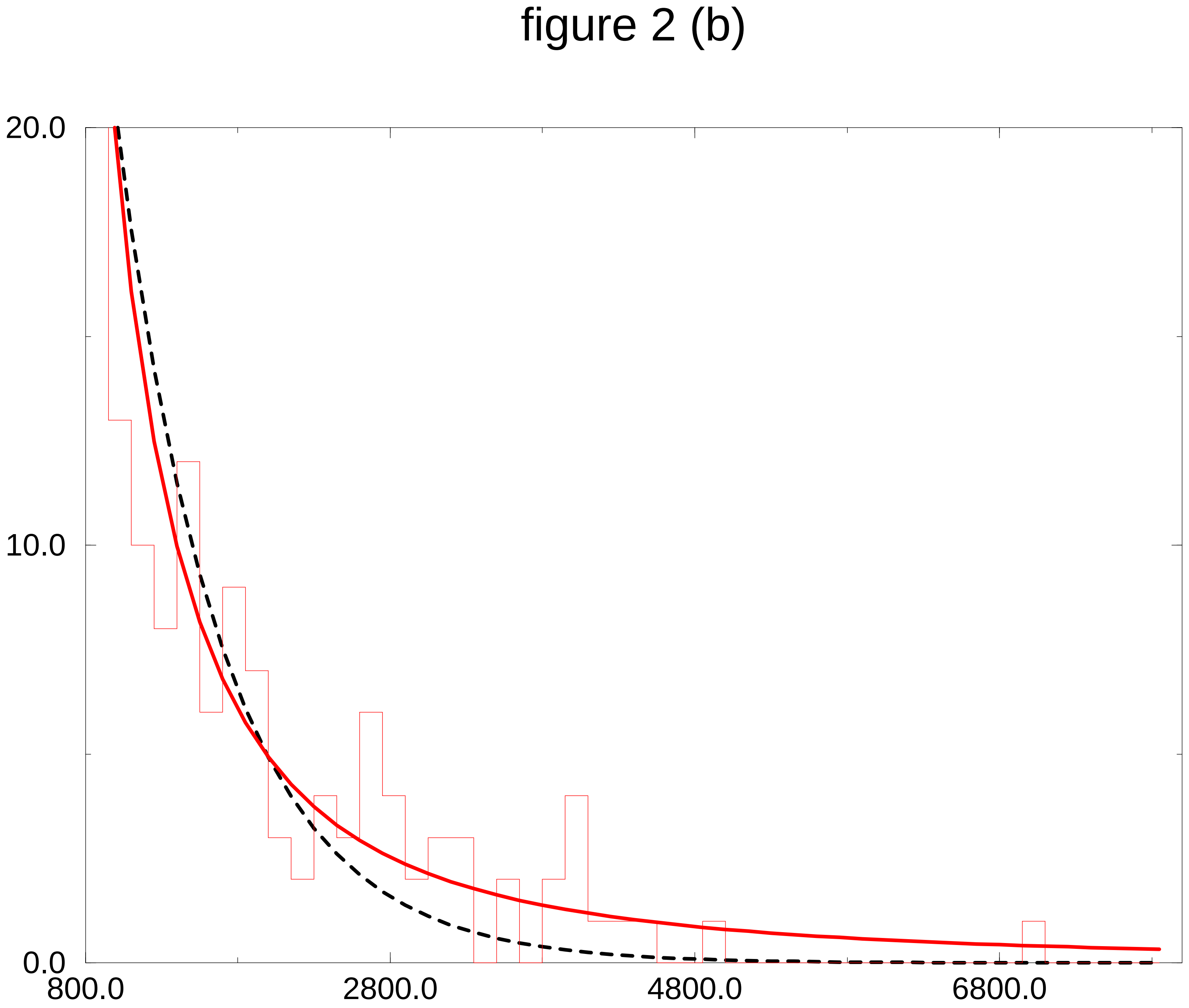


figure 2 (b)
20.0
10.0
0.0
800.0
2800.0
4800.0
6800.0

figure 2 (c)

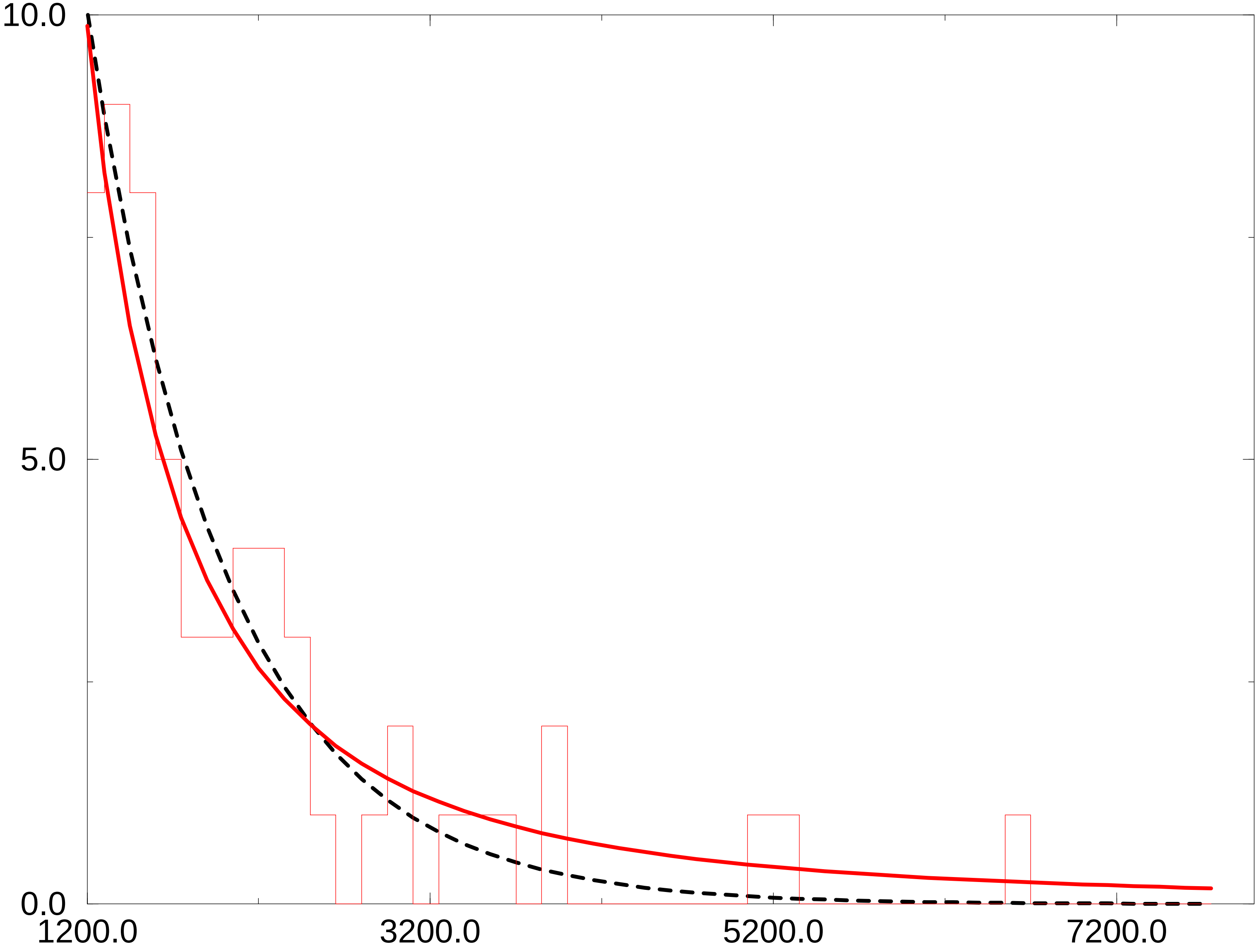


Figure 3 (a)

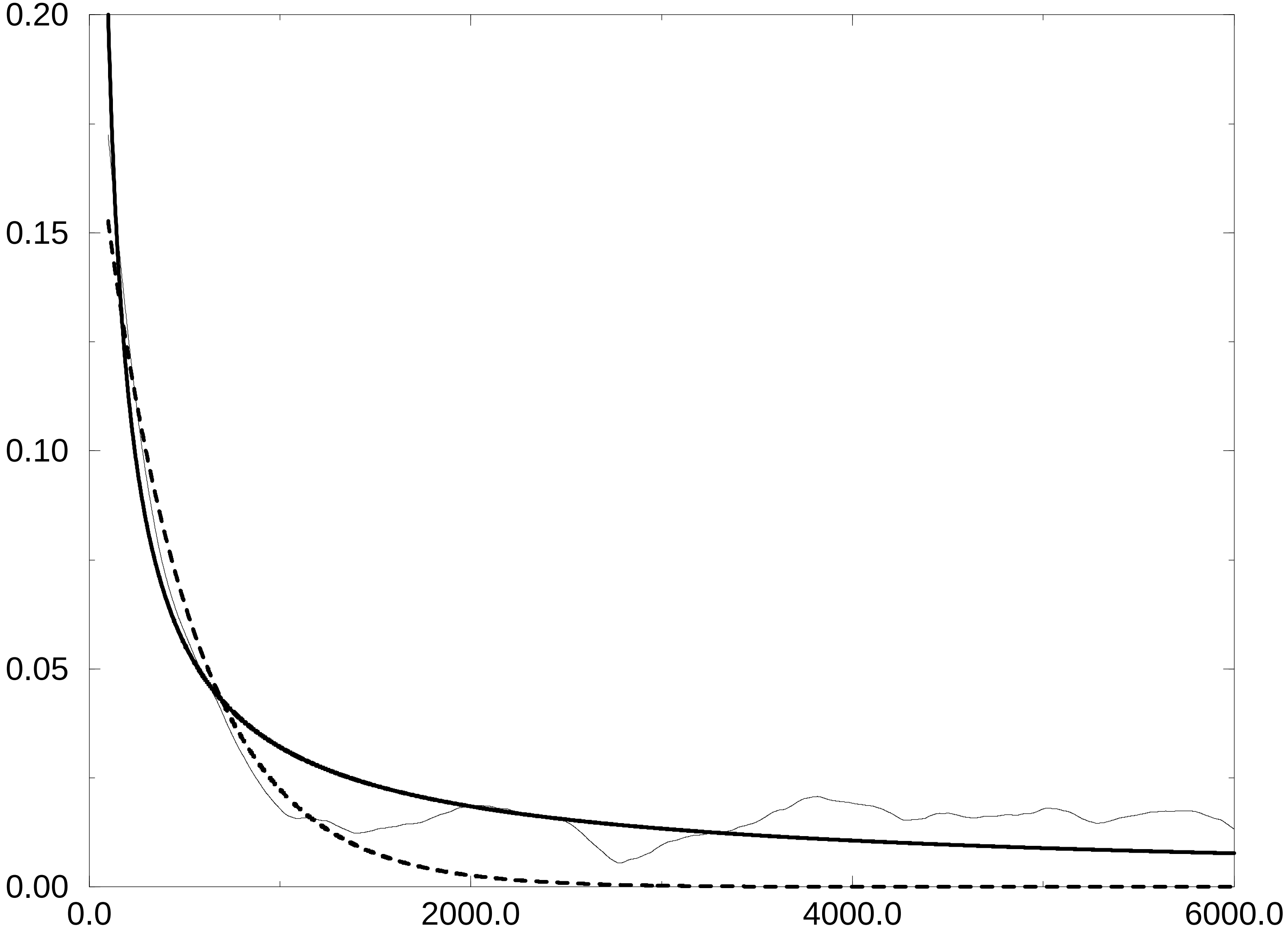


Figure 3 (b)

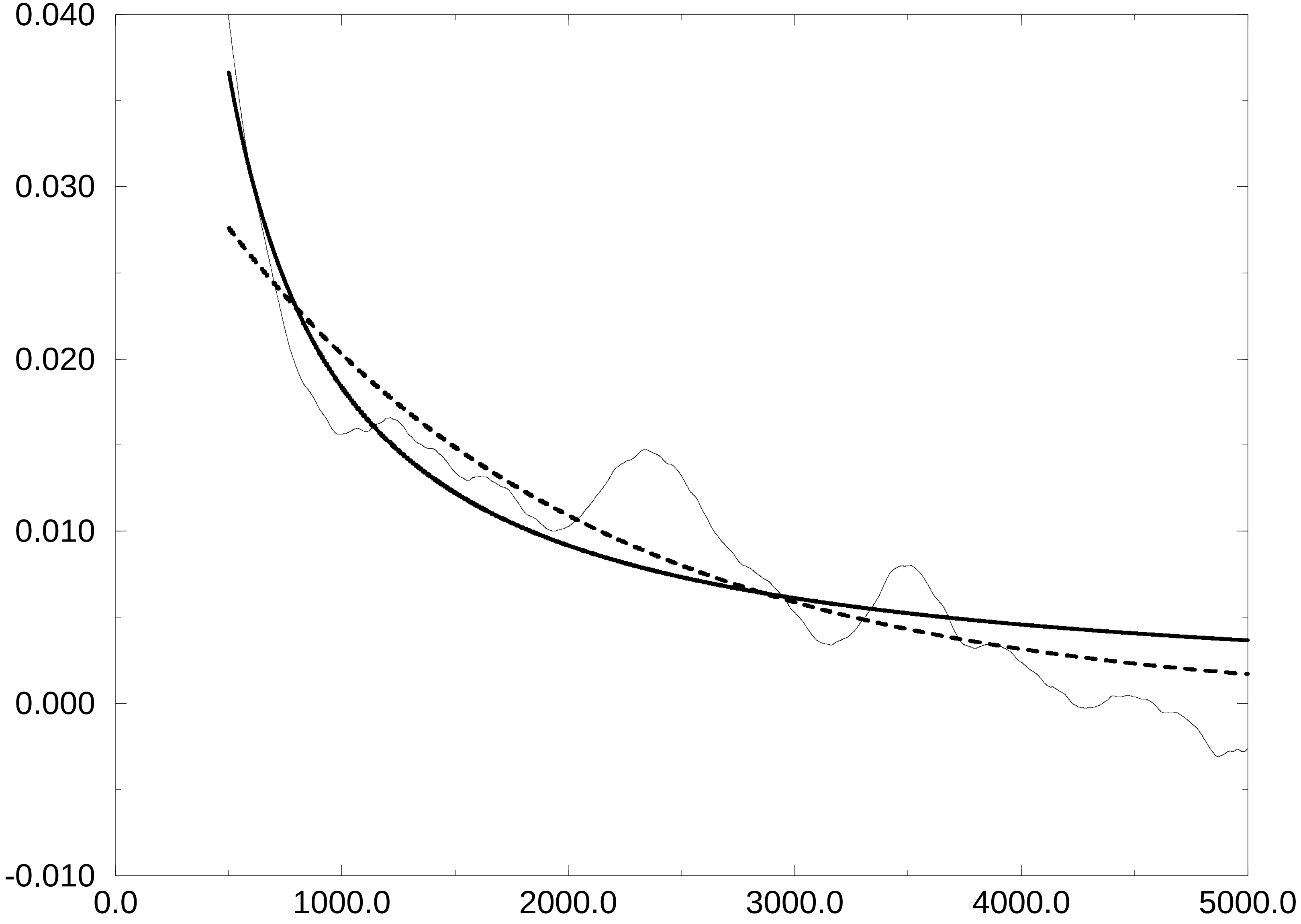
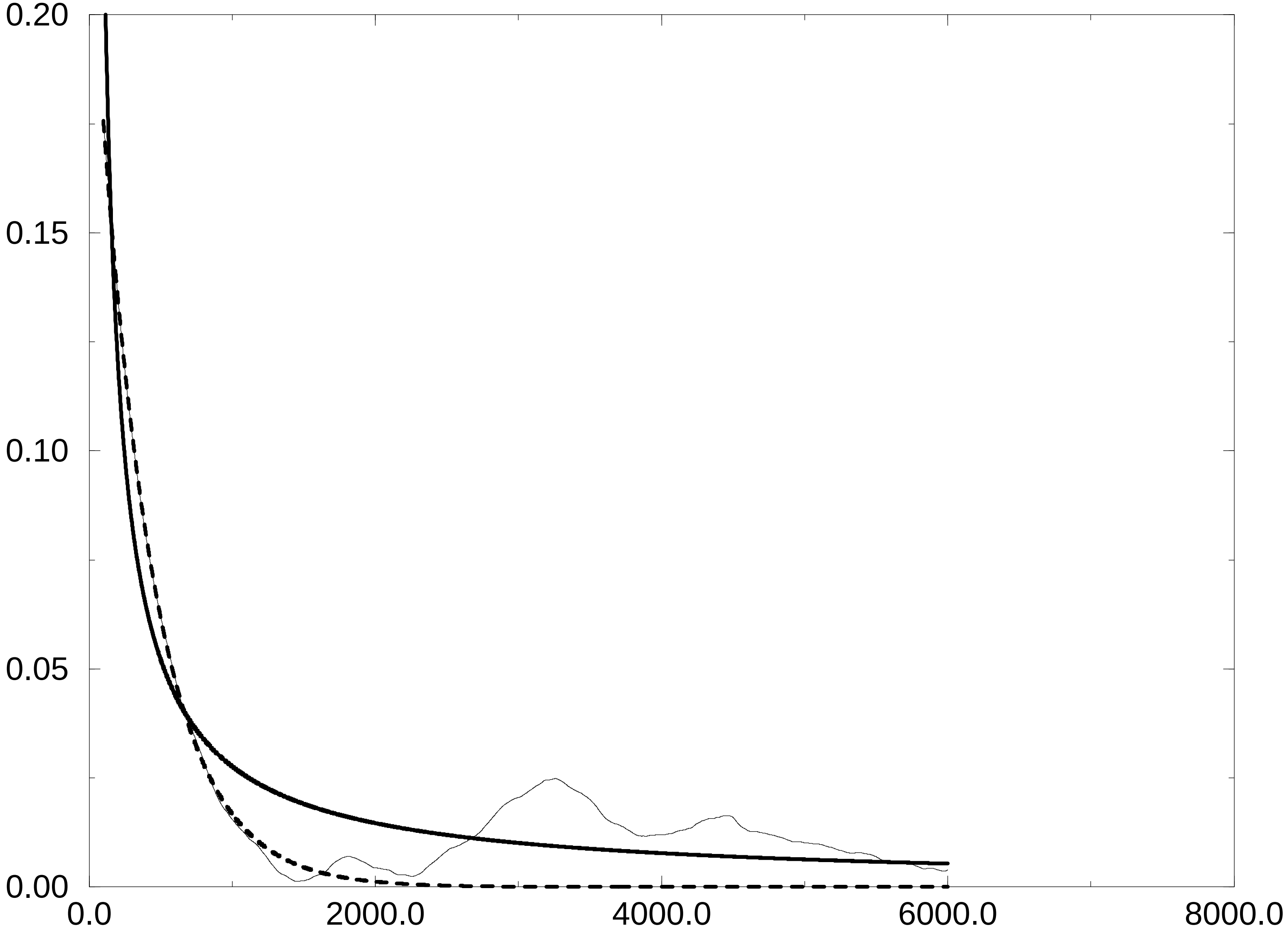

0.040
0.030
0.020
0.010
0.000
-0.010
0.0
1000.0
2000.0
3000.0
4000.0
5000.0

Figure 3 (c)


0.20
0.15
0.10
0.05
0.00
0.0
2000.0
4000.0
6000.0
8000.0